# Diagnosing Subword Boundary Conflicts in Diffusion Language Models through Sentence Simplification

**Yaowen Xu**[1], **Ming Cai**[2]

[1]College of Computer Science and Technology, [2]College of Artificial Intelligence
Zhejiang University, Hangzhou, China

**Correspondence:** cm@zju.edu.cn

## Abstract

Sentence simplification is a source-grounded rewriting task in which outputs often keep parts of the source while editing adjacent spans. Diffusion language models (DLMs) are well suited to this setting because masked denoising can fill rewritten spans using bidirectional context. However, parallel filling can create subword-boundary errors when incompatible subword fragments are committed together. We define these errors as *Subword Boundary Conflicts* (SBCs): malformed output words, such as *Australianways*, formed from incompatible fragments at subword boundaries. To measure them, we introduce a surface-form taxonomy and a detection pipeline based on candidate extraction, LLM-as-a-Judge filtering, and human verification. We evaluate Dream and LLaDA variants on ASSET and MED-EASi, with autoregressive and task-specific simplification baselines. We find that SBCs occur across DLMs but vary by model. Under standard unmasking, LLaDA variants remain at or below 2.80%, whereas Dream-7B-Instruct reaches 6.32% on ASSET and 9.03% on MED-EASi. Beyond model choice, decoding parallelism is a key driver. In the Dream unmasking setting at 128 denoising steps, increasing parallelism from 1 to 4 raises SBC rates from near zero to 25.63% on ASSET and 34.67% on MED-EASi. Further analyses show that SBCs occur under light-to-moderate edits, are not limited to rare words, and can often be repaired by local masked span regeneration. These findings establish SBCs as a word-form reliability target for robust diffusion-based text generation.

## 1 Introduction

Sentence simplification rewrites complex sentences into more accessible text while preserving the source meaning (Saggion and Hirst, 2017; Alva-Manchego et al., 2020b). It benefits readers such as children, non-native speakers, and laypeople who need to understand technical or medical information (Devaraj et al., 2022; Wu et al., 2023, 2024; Li et al., 2025a). As a source-grounded rewriting task, sentence simplification requires systems to reduce lexical and syntactic complexity while preserving fluency, meaning, and the correct use of domain terminology (Xu et al., 2015, 2016; Agrawal and Carpuat, 2024; Trienes et al., 2024). This often involves local edit operations such as lexical replacement, modifier deletion, and copying source names, entities, or technical terms.

Diffusion language models (DLMs) are naturally suited for such local edits. By leveraging bidirectional context during masked denoising, they can reconstruct targeted spans while preserving the surrounding unedited text (Nie et al., 2025; Ye et al., 2025; Li et al., 2025b; Ren et al., 2026). However, the non-autoregressive parallel filling mechanism often used in these models can introduce subword-boundary errors. A model may copy part of a source word and generate an adjacent new word at the same time; if these adjacent tokens are committed too early, the fragments can be fused into invalid forms. This risk is especially visible in sentence simplification, where outputs often reuse or lightly edit source spans rather than being generated from scratch (Cardon and Bibal, 2023).

We define these errors as *Subword Boundary Conflicts* (SBCs), which are malformed output words caused by incompatible subword fragments at subword boundaries. Table 1 shows cases involving fragment fusions such as *Australian+ways* and *people+irium*. In *peopleirium*, the valid word *people* from the reference is fused with the suffixal fragment *irium* from *delirium*, making the error a boundary conflict rather than a simple misspelling. These errors are particularly problematic in medical simplification because malformed words can obscure technical concepts (Devaraj et al., 2022).

Existing automated simplification metrics do not directly measure SBCs. SARI, BLEU, and LENS primarily evaluate outputs at the sentence or



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 17196–17212
October 24-29, 2026 ©2026 Association for Computational Linguistics

| Dataset | Source | Sentence |
|---|---|---|
| ASSET | Complex | Before 1980, a section of the Commonwealth Railways Central Australian line passed along the western side of the Simpson Desert. |
| | Reference | Before 1980, a part of the Central Australian line ran along the western side of the Simpson Desert. |
| | Dream | Before 1980, a part of the Central Australianways line went through the Simpson Desert. |
| MED-EASi | Complex | In the most severe form (called malignant yellow fever), delirium, intractable hiccups, seizures, coma, and multiple organ failure may occur terminally. |
| | Reference | In the most severe form of yellow fever, people may experience delirium, persistent hiccups, seizures, coma, and multiple organ failure. |
| | Dream | In the most severe form of yellow fever, peopleirium, hiccups hiccups, seizures, coma, and multiple organ failure may occur. |

Table 1: Representative SBC examples. Highlighted fragments from the Complex or Reference sentences are erroneously fused into malformed words in the Dream (Dream-7B-Instruct) outputs.

edit level (Maddela et al., 2023; Heineman et al., 2023), so high overall metric scores do not rule out the presence of localized malformed word forms. We thus evaluate SBCs as a separate, fine-grained output-side reliability issue.

To accurately detect SBCs, one might consider off-the-shelf spell checkers. However, traditional spelling and grammatical error detection systems flag anomalous word forms merely as generic spelling errors (Kukich, 1992; Bryant et al., 2023). This is insufficient for SBCs because word-form-only detectors can overflag proper names and domain terms, and they can also fail to distinguish SBCs from ordinary spelling errors that do not arise from fragment fusion. Because identifying an SBC requires tracing the fragments of a malformed word back to the source or references, detection must inherently be context-aware.

Motivated by this context dependence, we define a comprehensive surface-form taxonomy and a two-stage context-based detection pipeline. The pipeline first uses a high-recall heuristic to extract error candidates, then rigorously filters them with an LLM-as-a-Judge (Zheng et al., 2023) that compares the source, references, and system outputs to trace whether a malformed form can be explained by fragment fusion. Applying this pipeline, we evaluate representative DLMs (Dream and LLaDA variants) against autoregressive and task-specific simplification baselines on the ASSET and MED-EASi benchmarks.

Our results show that SBCs occur across DLMs but vary substantially by model and decoding setting. Under standard unmasking, LLaDA variants remain at or below 2.80%, whereas Dream-7B-Instruct reaches verified SBC rates of 6.32% on ASSET and 9.03% on MED-EASi. Decoding parallelism is a key driver. In the Dream unmasking setting at 128 denoising steps, increasing parallelism from 1 to 4 raises SBC rates from near zero to 25.63% on ASSET and 34.67% on MED-EASi. Further analyses reveal that SBCs are fundamentally local boundary-level errors rather than full-sentence planning failures. They are not merely byproducts of heavy edit loads or rare words, yet can often be effectively repaired via local masked span regeneration.

Our main contributions are threefold:

- **Problem definition.** We formalize SBCs as word-form errors in diffusion-based simplification and introduce a six-type surface taxonomy that distinguishes them from spelling errors and acceptable word-form changes.
- **Decoding-factor diagnosis.** We identify parallel token commitment as a key driver of SBCs. Higher parallelism substantially increases SBC rates, whereas additional denoising steps provide limited mitigation.
- **Empirical analysis.** With the proposed two-stage context-based detection pipeline, we quantify human-verified SBCs on ASSET and MED-EASi using Dream and LLaDA variants. We show that SBCs can be missed by sentence-level metrics, occur even under light-to-moderate edits, are not merely rare-word errors, and can often be corrected through local span regeneration.

## 2 Problem Definition, Taxonomy, and Detection

### 2.1 Definition

We define a Subword Boundary Conflict (SBC) as a malformed output word form whose surface realization suggests that incompatible subword units

have been fused, repeated, truncated, or otherwise corrupted around subword boundaries.

Formally, let $x$ denote the source sentence, $y$ the system output, and $r$ optional reference(s). An SBC instance is a malformed word form $w \in y$, even when its fragments can be traced to words in $x$ or $r$. The error is defined on this surface form rather than at the sentence level.

Crucially, not every erroneous output is an SBC. We count a case as an SBC only when the malformed word suggests fragment-level corruption. As illustrated in Table 1, sentence-level fluency errors, such as whole-word repetitions where each repeated word is well-formed (e.g., *hiccups hiccups*), are excluded from this definition.

### 2.2 Taxonomy

We classify SBCs into six surface types: overlapping duplication (T1), prefix truncation (T2), suffix truncation (T3), internal deletion (T4), internal substitution (T5), and invalid concatenation (T6). Appendix E.1 provides sentence-level examples.

**T1: Overlapping Duplication Error.** A character sequence or subword fragment is repeated within the malformed word (e.g., *Chynayna* repeats the final fragment of *Chyna*).

**T2: Prefix Truncation Error.** The beginning of the intended word is missing, leaving a suffix-like residue (e.g., *ored* preserves only the ending of *colored*).

**T3: Suffix Truncation Error.** The ending of the intended word is missing, leaving a prefix-like residue (e.g., *trib* preserves only the beginning of *tributaries*).

**T4: Internal Deletion Error.** A middle fragment is missing while both outer edges remain (e.g., *Stehen* preserves the outer parts of *Stephen*).

**T5: Internal Substitution Error.** A middle fragment is replaced while the word's outer shape remains (e.g., *Venderina* keeps the outer fragments of *Vezina*).

**T6: Invalid Concatenation Error.** Separate words or stems are joined into one invalid surface form (e.g., *knownSandro* fuses parts that should remain separate).

This taxonomy is based on the visible malformed word rather than on the model's decoding trajectory. When multiple patterns apply, we assign the first to avoid double counting.

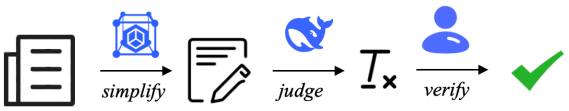



Figure 1: SBC detection pipeline. Candidate word forms absent from the source and references are filtered by `DeepSeek-V4-Pro` and manually verified in context.

### 2.3 Context-based Detection Pipeline

We detect SBCs in three steps. First, we collect output word forms absent from the source and references. Second, an LLM judge filters potential SBCs by tracing candidate fragments to the source or references, which helps distinguish boundary conflicts from typos or proper nouns. Finally, we manually verify the flagged cases. Figure 1 summarizes this pipeline.

**Candidate Extraction.** We split each final system output into surface word forms by whitespace after removing leading or trailing punctuation. We then keep word forms that do not appear in either the source $x$ or the references $r$ under case-insensitive matching. This candidate step is intentionally broad, so it includes malformed words such as *Australianways* as well as valid output words that are removed in subsequent steps.

**LLM-as-a-Judge Filtering.** We prompt `DeepSeek-V4-Pro` with the source, references, output, and candidate list. The judge is restricted to flagging candidates whose surface form suggests fragment-level corruption (e.g., invalid concatenation, duplication, truncation). Absence from the source/references alone is not evidence of an error. The judge returns a binary decision, the flagged forms, and a rationale. See Appendix A.2 for the prompt and Appendix B.2 for inference settings and auxiliary results.

**Human Verification.** Two authors with NLP research experience independently reviewed every judge-positive candidate against the source, references, and system output, labeling it as an SBC or non-SBC. Initial disagreements were resolved through discussion, and only consensus-positive cases were included in the reported SBC counts.

## 3 Experimental Setup

### 3.1 Datasets

Our main experiments use two representative English sentence simplification benchmarks.

**ASSET** (Alva-Manchego et al., 2020a) is a multi-reference Wikipedia corpus where each source sentence is paired with ten human-written references. By encompassing diverse edit operations, such as lexical replacement, sentence splitting, and content deletion, ASSET serves as a comprehensive benchmark for general-domain sentence simplification.

**MED-EASi** (Basu et al., 2023) consists of clinical sentences written for medical professionals and their plain-language versions for patients or other non-expert readers. Unlike general-domain datasets, MED-EASi often rewrites medical terminology through paraphrase, expansion, and brief definitions, producing longer targets. These terminology-heavy edits create frequent decisions about whether rare word forms should be preserved, split, or rewritten, making MED-EASi a targeted stress test for SBC detection.

For overall simplification quality, we follow prior work and evaluate on the official test splits: 359 ASSET examples and 300 MED-EASi examples. For DLM-specific SBC analysis, we additionally include ASSET's official validation split and MED-EASi's training and validation splits, yielding 2,359 ASSET examples and 1,893 MED-EASi examples. This larger pool supports more stable diagnostic analysis.

### 3.2 Models

Our main experiments evaluate three complementary model families. Our primary focus is on DLMs, including Dream-7B-Instruct (Ye et al., 2025), LLaDA-1.5 (Zhu et al., 2025), and LLaDA-8B-Instruct (Nie et al., 2025). For autoregressive (AR) comparisons, we use Qwen2.5-7B-Instruct (Yang et al., 2024), the base model for Dream-7B-Instruct, and LLaMA-3-8B-Instruct (Grattafiori et al., 2024), the AR reference model for LLaDA. We also include two task-specific simplification baselines, BART-based MUSS (Martin et al., 2022) and T5-based Control-T5 (Sheang and Saggion, 2021).

For the DLMs, we compare three decoding strategies with different commitment policies: standard unmasking, random filling, and remasking. Here, commitment means replacing a masked position with a model prediction. Standard unmasking selects positions by entropy for Dream or confidence for LLaDA, while random filling selects them randomly; both keep committed predictions fixed. For remasking, we apply ReMDM (Wang et al., 2025) to both model families, revisiting low-confidence predictions late in decoding.

All general-purpose LLMs use the same simplification instruction adapted from BLESS (Kew et al., 2023). The prompt template is provided in Appendix A.1. Full inference settings are provided in Appendix B.1.

### 3.3 Metrics

We evaluate overall simplification quality using SARI (Xu et al., 2016), BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019), and LENS (Maddela et al., 2023). These metrics are computed on the test splits with all available references, ten for ASSET and one for MED-EASi; SARI and BLEU follow the EASSE setup (Alva-Manchego et al., 2019).

For SBC analysis, we report the number and percentage of outputs containing at least one human-verified SBC as **error count** and **error rate**, respectively.

## 4 Experimental Results

We first report overall simplification quality and verified SBC rates, then analyze how decoding parallelism, denoising steps, editing load, and lexical frequency shape these boundary failures. We then examine attention patterns and local regeneration, and finally test generality across another source-grounded editing task and a block-diffusion architecture.

### 4.1 Main Results

**DLMs are competitive simplifiers, but decoding matters.** Table 2 shows that DLMs can be competitive on aggregate metrics but do not consistently outperform task-specific and autoregressive systems. Their behavior is strongly shaped by how tokens are committed during decoding. Dream-7B-Instruct attains the strongest ASSET BLEU under remasking, with a near-identical score under unmasking (74.65 and 74.41, respectively), whereas random filling sharply reduces BLEU and $\text{BERT}_{F1}$ (BLEU $74.41 \rightarrow 63.23$; $\text{BERT}_{F1}$ $76.54 \rightarrow 68.70$). By contrast, both LLaDA variants benefit from random filling on most metrics, with the effect strongest for LLaDA-8B-Instruct, which improves across all metrics on both datasets, including ASSET SARI ($42.61 \rightarrow 47.20$) and $\text{BERT}_{F1}$ ($73.47 \rightarrow 79.84$). This contrast suggests that the interaction between model-specific denois-

| Family | Model | Method | ASSET | | | | MED-EASi | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | SARI | BLEU | $\text{BERT}_{F1}$ | LENS | SARI | BLEU | $\text{BERT}_{F1}$ | LENS |
| Reference | Human references | – | 44.89 | 68.95 | 78.59 | 65.13 | – | – | – | – |
| Task-specific | MUSS | – | 44.90 | 68.29 | 78.89 | 69.23 | 35.12 | 21.07 | **44.55** | 59.16 |
| Task-specific | Control-T5 | – | 45.08 | 71.03 | **80.59** | 66.29 | 35.64 | **21.95** | 44.16 | 55.45 |
| Autoregressive | LLaMA-3-8B-Instruct | – | 46.19 | 50.11 | 75.09 | 69.77 | 40.10 | 10.96 | 39.18 | **66.32** |
| Autoregressive | Qwen2.5-7B-Instruct | – | **48.17** | 58.72 | 79.71 | **69.96** | **40.63** | 13.48 | 43.66 | 62.85 |
| DLM | Dream-7B-Instruct | unmasking | 45.20 | 74.41 | 76.54 | 62.43 | 39.00 | 15.57 | 39.25 | 50.59 |
| DLM | Dream-7B-Instruct | remasking | 45.22 | **74.65** | 76.67 | 62.49 | 39.00 | 15.56 | 39.33 | 51.42 |
| DLM | Dream-7B-Instruct | random filling | 42.76 | 63.23 | 68.70 | 63.31 | 38.43 | 9.34 | 29.46 | 56.50 |
| DLM | LLaDA-1.5 | unmasking | 45.66 | 63.83 | 78.44 | 68.31 | 37.33 | 8.27 | 37.27 | 58.09 |
| DLM | LLaDA-1.5 | remasking | 45.72 | 64.13 | 78.45 | 68.26 | 37.32 | 8.27 | 37.29 | 58.62 |
| DLM | LLaDA-1.5 | random filling | 47.09 | 60.48 | 78.79 | 69.16 | 39.69 | 12.51 | 42.53 | 61.67 |
| DLM | LLaDA-8B-Instruct | unmasking | 42.61 | 58.49 | 73.47 | 66.65 | 35.96 | 4.24 | 31.54 | 53.06 |
| DLM | LLaDA-8B-Instruct | remasking | 42.65 | 58.61 | 73.50 | 66.66 | 35.94 | 4.23 | 31.53 | 53.06 |
| DLM | LLaDA-8B-Instruct | random filling | 47.20 | 63.51 | 79.84 | 69.86 | 39.78 | 12.74 | 42.48 | 60.49 |

Table 2: Automatic evaluation of simplification quality on the ASSET and MED-EASi test sets. $\text{BERT}_{F1}$ scores are scaled to $[0, 100]$; higher values indicate better performance for all metrics. Best scores in each metric column are boldfaced, and the best DLM scores are underlined.

| Model | Method | ASSET | | | MED-EASi | | | |
|---|---|---|---|---|---|---|---|---|
| | | Test (359) | Valid (2000) | All (2359) | Test (300) | Valid (196) | Train (1397) | All (1893) |
| Dream-7B-Instruct | unmasking | 19 (5.29) | 130 (6.50) | 149 (6.32) | 33 (11.00) | 20 (10.20) | 118 (8.45) | 171 (9.03) |
| Dream-7B-Instruct | remasking | 20 (5.57) | 141 (7.05) | 161 (6.82) | 35 (11.67) | 26 (13.27) | 105 (7.52) | 166 (8.77) |
| Dream-7B-Instruct | random filling | 4 (1.11) | 20 (1.00) | 24 (1.02) | 1 (0.33) | 7 (3.57) | 26 (1.86) | 34 (1.80) |
| LLaDA-1.5 | unmasking | 5 (1.39) | 35 (1.75) | 40 (1.70) | 8 (2.67) | 9 (4.59) | 36 (2.58) | 53 (2.80) |
| LLaDA-1.5 | remasking | 4 (1.11) | 31 (1.55) | 35 (1.48) | 7 (2.33) | 7 (3.57) | 36 (2.58) | 50 (2.64) |
| LLaDA-1.5 | random filling | 1 (0.28) | 7 (0.35) | 8 (0.34) | 3 (1.00) | 5 (2.55) | 16 (1.15) | 24 (1.27) |
| LLaDA-8B-Instruct | unmasking | 4 (1.11) | 28 (1.40) | 32 (1.36) | 5 (1.67) | 3 (1.53) | 24 (1.72) | 32 (1.69) |
| LLaDA-8B-Instruct | remasking | 5 (1.39) | 30 (1.50) | 35 (1.48) | 5 (1.67) | 3 (1.53) | 26 (1.86) | 34 (1.80) |
| LLaDA-8B-Instruct | random filling | 1 (0.28) | 6 (0.30) | 7 (0.30) | 2 (0.67) | 5 (2.55) | 11 (0.79) | 18 (0.95) |

Table 3: Human-verified prevalence of Subword Boundary Conflicts (SBCs) across diagnostic splits for diffusion language models. Counts are shown with split-level rates in parentheses.

ing behavior and token-commitment strategy warrants investigation. Yet these metrics surface only sentence-level behavior, leaving open whether the same token-commitment strategies preserve subword boundary integrity at the word-form level.

**SBCs are decoding-sensitive but invisible to standard metrics.** Table 3 highlights three trends. First, under standard unmasking/remasking, SBC rates vary widely: Dream-7B-Instruct reaches 6.32%–9.03%, while LLaDA variants stay under 2.80%. Second, random filling sharply reduces SBCs, particularly for Dream (dropping from $\sim$9% to $\sim$1.8% on MED-EASi). Third, this reduction is not consistently captured by sentence-level quality. The same decoding change drops Dream's BLEU by over 11 points (Table 2), while LLaDA-8B-Instruct improves across all metrics. Standard sentence metrics cannot proxy SBCs, establishing them as an independent reliability dimension.

**SBC surface types vary across models and decoding strategies.** Table 4 shows diverse surface patterns. Under unmasking and remasking, Dream errors concentrate in overlapping duplication (T1) and invalid concatenation (T6), whereas LLaDA favors suffix truncation (T3) and T6. Within a model, unmasking and remasking yield similar profiles, indicating the current remasking schedule does not mitigate SBCs. By contrast, random filling reduces total counts and alters the type distribution. Overall, invalid concatenation (T6) recurs across DLMs, while T1 and T3 reveal model-specific failure patterns hidden by aggregate SBC rates.

### 4.2 Parallelism and Denoising Steps

We run two controlled ablations under standard unmasking to examine how parallelism and denoising steps affect SBC rates. Dream-7B-Instruct is the main diagnostic setting because it is a representative high-SBC standard-unmasking case (Table 3). We vary parallelism at 128 steps and vary denoising steps at parallelism 2. Appendix C mirrors these ablations for LLaDA-1.5.

**Higher Parallelism Increases Boundary Conflicts.** At 128 steps, SBC rates rise from near zero at parallelism 1 to 25.63% on ASSET and 34.67% on MED-EASi at parallelism 4, remaining high at parallelism 8 (Table 5). Although aggregate rates plateau, the specific errors shift: many are shared between parallelisms 4 and 8, but each introduces $\sim$30 unique cases per dataset.

| Model | Method | ASSET | | | | | | MED-EASi | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T1 | T2 | T3 | T4 | T5 | T6 | T1 | T2 | T3 | T4 | T5 | T6 |
| Dream-7B-Instruct | unmasking | 52 | 17 | 26 | 6 | 5 | 51 | 58 | 20 | 37 | 10 | 10 | 47 |
| Dream-7B-Instruct | remasking | 59 | 29 | 23 | 6 | 6 | 46 | 63 | 26 | 34 | 9 | 10 | 41 |
| Dream-7B-Instruct | random filling | 4 | 3 | 7 | 8 | 1 | 1 | 12 | 1 | 4 | 13 | 0 | 4 |
| LLaDA-1.5 | unmasking | 11 | 0 | 12 | 4 | 3 | 12 | 9 | 2 | 12 | 8 | 2 | 23 |
| LLaDA-1.5 | remasking | 10 | 0 | 8 | 4 | 2 | 12 | 8 | 3 | 12 | 7 | 2 | 19 |
| LLaDA-1.5 | random filling | 1 | 2 | 3 | 1 | 1 | 1 | 0 | 2 | 13 | 6 | 1 | 6 |
| LLaDA-8B-Instruct | unmasking | 6 | 0 | 6 | 3 | 3 | 14 | 4 | 1 | 13 | 1 | 1 | 14 |
| LLaDA-8B-Instruct | remasking | 8 | 0 | 6 | 4 | 1 | 16 | 3 | 2 | 13 | 2 | 1 | 13 |
| LLaDA-8B-Instruct | random filling | 2 | 0 | 0 | 0 | 0 | 5 | 0 | 1 | 3 | 12 | 0 | 5 |

Table 4: Distribution of Subword Boundary Conflict (SBC) surface types, reported as counts of human-verified malformed word forms. Row totals may exceed the instance-level counts in Table 3 because a single output can contain multiple malformed word forms. T1: overlapping duplication; T2: prefix truncation; T3: suffix truncation; T4: internal deletion; T5: internal substitution; T6: invalid concatenation.

| Parallelism | Steps | ASSET Test (359) | MED-EASi Test (300) |
|---|---|---|---|
| 1 | 128 | 1 (0.28) | 2 (0.67) |
| 2 | 128 | 19 (5.29) | 33 (11.00) |
| 4 | 128 | 92 (25.63) | 104 (34.67) |
| 8 | 128 | 95 (26.46) | 98 (32.67) |

Table 5: SBC rates for Dream-7B-Instruct standard unmasking outputs across parallelism values, with the denoising budget fixed at 128 steps.

| Parallelism | Steps | ASSET Test (359) | MED-EASi Test (300) |
|---|---|---|---|
| 2 | 32 | 27 (7.52) | 36 (12.00) |
| 2 | 64 | 20 (5.57) | 30 (10.00) |
| 2 | 128 | 19 (5.29) | 33 (11.00) |
| 2 | 256 | 24 (6.69) | 26 (8.67) |
| 2 | 512 | 14 (3.90) | 28 (9.33) |

Table 6: SBC rates for Dream-7B-Instruct standard unmasking outputs across denoising-step budgets, with parallelism fixed at 2.

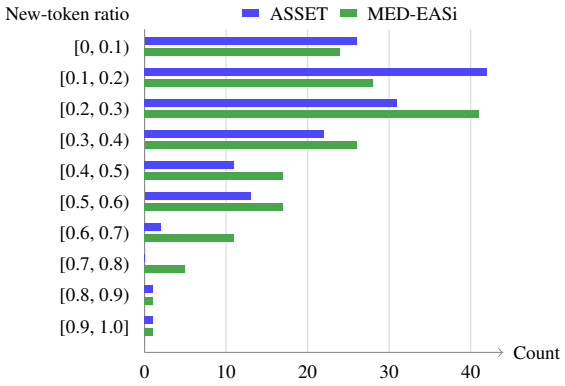



Figure 2: Distribution of human-verified Dream-7B-Instruct unmasking SBCs by new-token ratio.

**Increasing Denoising Steps Does Not Consistently Reduce Boundary Conflicts.** Across the tested range from 32 to 512 steps, SBC rates do not decrease steadily. ASSET reaches 5.29% at 128 steps, rises to 6.69% at 256 steps, and then falls to 3.90% at 512 steps, while MED-EASi remains between 8.67% and 12.00% (Table 6). Since ASSET is non-monotonic and MED-EASi stays above 8.67% throughout, increasing the step budget alone does not consistently reduce boundary conflicts.

The corresponding LLaDA-1.5 ablations show the same pattern (Appendix C), further indicating that per-iteration token commitment, rather than an insufficient denoising budget, is a key driver of boundary conflicts.

### 4.3 Editing Load

**Do SBCs mainly reflect extensive rewriting?** To answer this, we quantify editing load with the *new-token ratio*: the proportion of output tokens not copied from the source. After aligning the tokenized source and output sequences, we treat any output token introduced by an insertion or replacement as "new." This token-level measure more directly reflects the amount of generation required from the model than word-level edit distance. For example, editing $\langle A, B, C, D\rangle$ to $\langle A, B, X, D, Y\rangle$ produces two new tokens $(X, Y)$ out of five output tokens, yielding a ratio of 0.4. We compute this ratio over all human-verified Dream-7B-Instruct unmasking SBC cases (Figure 2). If extensive rewriting were the main driver, these errors would concentrate in high-ratio bins.

**SBCs mainly occur under light-to-moderate editing.** Most verified errors fall below a new-token ratio of 0.4: 81.2% (121/149) for ASSET and 69.6% (119/171) for MED-EASi. The distributions peak early, though MED-EASi shows a slightly larger high-editing tail (10.5% of cases at ratios $\geq 0.6$, vs. 2.7% for ASSET) due to terminology expansions. However, this tail is not the dominant pattern. Thus, SBCs frequently arise even when the model preserves most of the source, confirming they are localized boundary failures rather than

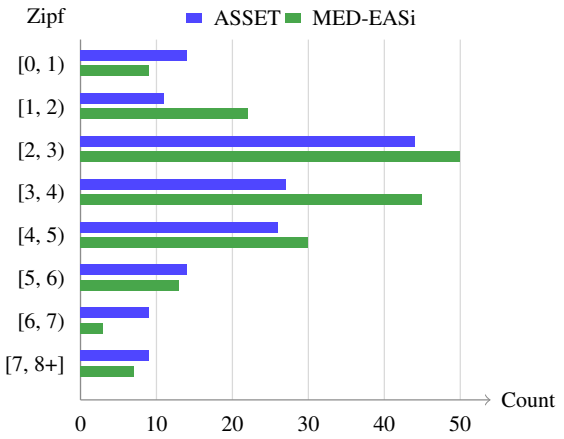



Figure 3: Zipf profile of matched source-word observations for verified Dream-7B-Instruct unmasking SBCs.

byproducts of extensive rewriting.

### 4.4 Word Frequency

**Are SBCs strictly a rare-word phenomenon?** We measure lexical frequency by matching each verified unmasking SBC to its source word(s) and computing their Zipf frequencies. Since an SBC can map to multiple source words, we use matched source-word observations rather than instance-level cases (154 for ASSET, 179 for MED-EASi). Zipf frequency is a log-scaled estimate where higher values denote more common words (e.g., Zipf 3 is about once per million; Zipf 6 is about once per thousand) [1]. If rare-vocabulary unfamiliarity were the main driver, errors would cluster in the lowest-frequency bins.

**SBCs are not limited to rare words.** While SBCs skew toward lower-frequency source words, they are not restricted to rare vocabulary (Figure 3). Most observations fall below Zipf 4 (62.3% for ASSET, 70.4% for MED-EASi), peaking in the $[2, 3)$ bin. Yet errors persist for frequent words (Zipf $\geq 5$), comprising 20.8% and 12.8% of ASSET and MED-EASi observations, respectively. Thus, low frequency is not a necessary condition for SBCs. For frequent words, boundary conflicts likely stem from abrupt local context shifts (e.g., localized substitution) that force otherwise familiar subwords into invalid fusion during parallel generation.

### 4.5 Attention

This case illustrates an SBC containing a subword also found in the source. In Table 1 (ASSET), the source contains *Commonwealth Railways Central Australian line*. While the reference retains *Central Australian line*, Dream produces the malformed *Central Australianways line*, fusing the preserved word *Australian* with the subword *ways* from *Railways*. The shift-aware attention visualization in Appendix E.4 shows an attention pattern consistent with local copying or alignment behavior. At decoding step 6, the query predicting the erroneous *ways* subword attends to the source-side *ways* subword in *Railways*. At step 7, after the token sequence underlying *Australianways* has entered the decoded context, the query predicting *line* assigns more attention to the output-side *ways* token than to the source-side *line*. These observations suggest that this SBC is less likely to reflect a failure of semantic understanding than a local copying or alignment issue during parallel decoding. A fuller mechanistic account remains for future work.

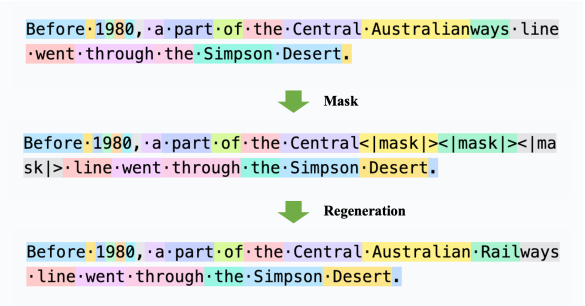



Figure 4: Mask-guided local regeneration for the *Australianways* case. Local refilling replaces the malformed span with *Australian Railways* while the rest of the output remains fixed.

### 4.6 Mask-Guided Regeneration

Finally, we test whether SBCs can be corrected without regenerating the full sentence. We use this intervention to assess local recoverability by testing if the model can repair the malformed boundary with the surrounding sentence and source context fixed. Given the source $x$, output $y$, and a human-verified SBC word form, we replace its corresponding token span with mask tokens, appending one additional mask token to allow span expansion, e.g., *Australianways* $\rightarrow$ *Australian Railways* (Figure 4). Regeneration is performed sequentially to isolate local recoverability from parallel decoding failures: at each iteration, the model commits only the highest-confidence token among the remaining masked positions until all masks are filled. Representative unmasking cases across models and datasets appear in Table 16.

[1] https://github.com/rspeer/wordfreq

This post-hoc procedure assumes an identified malformed span and therefore serves only as a diagnostic probe, not a general correction method. We apply it to all human-verified SBC word forms from each model's default-unmasking test outputs and count a repair as successful only when the malformed form is replaced by a locally appropriate, well-formed word or phrase. On ASSET and MED-EASi, respectively, regeneration repaired 15/20 and 30/37 Dream-7B-Instruct word forms, 5/5 and 10/10 LLaDA-1.5 forms, and 4/4 and 4/5 LLaDA-8B-Instruct forms, yielding 68/81 (84.0%) successful repairs overall. These results show that many SBCs are locally recoverable without full-sentence replanning.

### 4.7 Generality Across Tasks and Architectures

We conduct two auxiliary evaluations to test whether SBCs extend beyond the main setting, with full results in Appendix D.

On JFLEG's 747-example grammatical error correction test set (Napoles et al., 2017), Dream-7B-Instruct and LLaDA-1.5 achieve GLEU scores of 61.12 and 58.77, respectively, compared with the published 40.54 for the uncorrected source (Coyne et al., 2023); each produces 7 outputs containing human-verified SBCs (0.94%). These nonzero rates show that SBCs also occur in another source-grounded editing task.

We further evaluate Fast-dLLM v2-7B, a block-diffusion model (Wu et al., 2026a). As its sub-block size increases from 4 to 32, SBC rates rise from 6.13% to 13.65% on ASSET and from 10.00% to 13.33% before plateauing on MED-EASi. These results extend the evidence to another task and DLM architecture, while remaining limited to English.

## 5 Discussion

Our results identify SBCs as a word-form reliability issue that is not captured by standard simplification metrics. Dream-7B-Instruct performs competitively on standard metrics while producing more verified SBCs than LLaDA variants. Random filling reduces SBCs for Dream, but it does so at the cost of sentence-level quality. SBCs should therefore be measured separately from standard simplification quality. Cases such as *Australianways* and *peopleirium* illustrate how SBCs may obscure names or clinically relevant terms, but we do not directly measure their effects on factuality, comprehension, or user trust.

We also find a few SBC-like errors in non-DLM outputs. For example, autoregressive systems sometimes attach quotation marks to neighboring words or start an output with suffix fragments such as *ly* and *ed*. Control-T5 occasionally produces fused or malformed forms such as *PakistanAdministrative* and *tighture* (Table 15). These cases require further study.

Mitigating SBCs may benefit from interventions during both decoding and training. During decoding, models could delay uncertain subword commitments, remask suspicious boundaries, or regenerate local spans. While our tests show many SBCs are locally fixable, this regeneration step is only a proof of concept. During training, objectives that reconstruct whole words or entities may enforce boundary consistency (He et al., 2026). Future work should disentangle the overlapping effects of tokenization, architecture, training objectives, and decoding order.

## 6 Related Work

**Sentence Simplification** Sentence simplification has evolved from linguistically motivated rewriting and monolingual translation to neural, edit-based, and controllable generation (Chandrasekar et al., 1996; Zhu et al., 2010; Zhang and Lapata, 2017; Dong et al., 2019; Sheang and Saggion, 2021; Martin et al., 2022). Benchmarks and metrics mainly assess simplicity, fluency, and faithfulness at the sentence or edit level (Alva-Manchego et al., 2019; Maddela et al., 2023; Heineman et al., 2023; Devaraj et al., 2022). Recent evaluations extend this scope to LLMs, readability control, multilinguality, and domain shift (Kew et al., 2023; Barayan et al., 2025; Naous et al., 2024; Scalercio et al., 2025), but still emphasize overall output quality. We instead use simplification as a source-grounded rewriting task to diagnose a fine-grained output word-form failure mode that conventional simplification metrics fail to capture.

**Diffusion Language Models** DLMs span continuous and discrete formulations, including masked variants that operate directly over token sequences (Austin et al., 2021; Li et al., 2022; Lou et al., 2023; Sahoo et al., 2024; Li et al., 2025b). Prior work on sentence simplification has evaluated sequence-to-sequence diffusion models, including DiffuSeq, TESS, and SeqDiffuSeq (Gong et al., 2023; Mahabadi et al., 2024; Yuan et al., 2024). Recent

large-scale masked DLMs show that iterative denoising with parallel token revision can serve as an alternative to left-to-right autoregressive decoding (Nie et al., 2025; Zhu et al., 2025; Ye et al., 2025). A central thread of recent work targets scalable and efficient DLM inference through hybrid generation, block diffusion, fast parallel decoding, and inference-time remasking (Wu et al., 2026a,b; Wang et al., 2025). Beyond efficiency, recent analyses have begun to characterize the broader consequences of parallel decoding, including its effects on speed, generation quality, semantic faithfulness, and coherence (Wu et al., 2026b; Kang et al., 2026). Using prompted masked DLMs (Dream and LLaDA), we complement recent parallel-decoding analyses by examining a finer-grained failure mode in which parallel denoising can construct malformed words from mutually incompatible subword fragments.

**Subword Boundary Conflicts** Tokenization work shows that boundary choices and non-canonical encodings affect model inputs and distributions (Zheng et al., 2025). SBCs are the output-side counterpart, consisting of malformed words caused by conflicts between adjacent subword fragments. We study them in source-grounded rewriting with DLMs.

## 7 Conclusion

This work identifies subword boundary conflicts (SBCs) as a word-form reliability issue in DLM-based sentence simplification. SBCs arise when incompatible subword fragments are committed at token boundaries, producing malformed surface forms. Using a surface-form taxonomy and LLM-as-a-Judge detection, our evaluations on ASSET and MED-EASi show SBCs occur systematically in DLMs. These errors are missed by standard metrics and increase with decoding parallelism; they occur even under light-to-moderate edits and are not limited to rare words. A masked-regeneration study further shows many SBCs are localized boundary errors rather than full-sentence failures. These findings establish SBCs as a concrete diagnostic target for improving diffusion-based text generation.

## Limitations

Our study is limited in three ways. First, our primary experiments focus on English sentence-level simplification and a limited set of DLM architectures. We do not evaluate document-level simplification, which may also be important for accessibility. Cross-lingual settings (Ryan et al., 2023) remain open directions. Second, our human verification focuses on judge-positive cases, and the LLM-as-a-Judge pipeline may still be inconsistent on ambiguous examples; broader domains, prompts, and exhaustive annotation would strengthen the estimates. Third, some of our analyses are diagnostic rather than strictly causal, and the local masked-regeneration study is a post-hoc intervention.

## Ethics Statement

This work follows the ACL Ethics Policy. We use publicly available corpora and models under their respective licenses, collect no new personal data, and do not deploy systems to end users. Because part of the analysis involves medical simplification, our findings should be read as model diagnostics rather than clinical guidance. LLM-as-a-Judge outputs may reflect model biases or inconsistencies, so we use them only as diagnostic evidence alongside human verification.

## References


Sweta Agrawal and Marine Carpuat. 2024. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020a. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. Easse: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020b. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-easi: Finely annotated dataset and models for controllable simplification of medical texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14093–14101.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, pages 643–701.

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–130.

Raman Chandrasekar, Christine Doran, and Srinivas Bangalore. 1996. Motivations and methods for text simplification. In *COLING 1996 volume 2: The 16th international conference on computational linguistics*.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

Ashwin Devaraj, William Sheffield, Byron C Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. Editnts: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. Diffuseq: Sequence to sequence text generation with diffusion models. In *The Eleventh International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shaokai He, Kaiwen Wei, Xinyi Zeng, Xiang Chen, Xue Yang, Zhenyang Li, Jiang Zhong, and Yu Tian. 2026. Differ: Diffusion entity-relation modeling for reversal curse in diffusion large language models. *arXiv preprint arXiv:2601.07347*.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing between success and failure: Edit-level simplification evaluation using salsa. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495.

Wonjun Kang, Kevin Galim, Seunghyuk Oh, Minjae Lee, Yuchen Zeng, Shuibai Zhang, Coleman Richard Charles Hooper, Yuezhou Hu, Hyung Il Koo, Nam Ik Cho, and Kangwook Lee. 2026. Parallelbench: Understanding the trade-offs of parallel decoding in diffusion LLMs. In *The Fourteenth International Conference on Learning Representations*.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. Bless: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM computing surveys (CSUR)*, 24(4):377–439.

Guanlin Li, Yuki Arase, and Noel Crespi. 2025a. Aligning sentence simplification with esl learner's proficiency for language acquisition. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 492–507.

Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen. 2025b. A survey on diffusion language models. *arXiv preprint arXiv:2508.10875*.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2023. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. Lens: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408.

Rabeeh Karimi Mahabadi, Hamish Ivison, Jaesung Tae, James Henderson, Iz Beltagy, Matthew E Peters, and Arman Cohan. 2024. TESS: Text-to-text self-conditioned simplex diffusion. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2347–2361.

Louis Martin, Angela Fan, Éric Villemonte De La Clergerie, Antoine Bordes, and Benoît Sagot. 2022. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 1651–1664.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. Readme++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. Jfleg: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Houxing Ren, Mingjie Zhan, Zimu Lu, Ke Wang, Yunqiao Yang, Haotian Hou, Junting Pan, and Hongsheng Li. 2026. Edit-based refinement for parallel masked diffusion language models. In *Forty-third International Conference on Machine Learning*.

Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-english text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927.

Horacio Saggion and Graeme Hirst. 2017. *Automatic text simplification*, volume 32. Springer.

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184.

Arthur Mariano Rocha De Azevedo Scalercio, Elvis A De Souza, Maria José B Finatto, and Aline Paes. 2025. Evaluating llms for portuguese sentence simplification with linguistic insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24452–24477.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352.

Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, Wei Xu, Byron C Wallace, and Junyi Jessy Li. 2024. Infolossqa: Characterizing and recovering information loss in text simplification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4263–4294.

Guanghan Wang, Yair Schiff, Subham Sahoo, and Volodymyr Kuleshov. 2025. Remasking discrete diffusion models with inference-time scaling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Shizhe Diao, Yonggan Fu, Zhijian Liu, Pavlo Molchanov, Ping Luo, Song Han, and Enze Xie. 2026a. Fast-dLLM v2: Efficient block-diffusion LLM. In *The Fourteenth International Conference on Learning Representations*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. 2026b. Fast-dLLM: Training-free acceleration of diffusion LLM by enabling KV cache and parallel decoding. In *The Fourteenth International Conference on Learning Representations*.

Guande Wu, Jing Qian, Sonia Castelo Quispe, Shaoyu Chen, João Rulff, and Claudio Silva. 2024. Artist: Automated text simplification for task guidance in augmented reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–24.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. 2024. Text diffusion model with encoder-decoder transformers for sequence-to-sequence generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 22–39.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.

Brian Siyuan Zheng, Alisa Liu, Orevaoghene Ahia, Jonathan Hayase, Yejin Choi, and Noah A. Smith. 2025. Broken tokens? your language model can secretly handle non-canonical tokenizations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and 1 others. 2025. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *arXiv preprint arXiv:2505.19223*.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.

# A Prompts

## A.1 Simplification Prompt

As detailed in Section 3, all general-purpose LLMs use the BLESS-adapted simplification prompt shown below (Kew et al., 2023).

**Simplification Prompt**

Please rewrite the following complex sentence in order to make it easier to understand by non-native speakers of English. You can do so by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified sentence needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning.
**Complex:** `{input_text}`
**Simple:**

## A.2 LLM-as-a-Judge Prompt

Figure 5 shows the LLM-as-a-Judge prompt used for SBC detection; full implementation details and inference settings are reported in Appendix B.2.

# B Inference Settings

## B.1 Model Inference Settings

Experiments use a workstation with four NVIDIA GeForce RTX 3090 GPUs and approximately 650 GPU hours. For reproducibility, all runs use seed 42 and the same evaluation splits. Simplification quality is evaluated on the full test split, and SBC diagnosis covers all splits described in Section 3.

Prompted autoregressive (AR) models and DLMs generate at most 256 new tokens. Prompted AR models are decoded using nucleus sampling with temperature $T = 1.0$ and top-$p$ threshold 0.9. For standard DLM unmasking, **Dream** uses entropy-based token scheduling with 128 denoising steps, temperature $T = 0.2$, and top-$p$ threshold 0.95. **LLaDA** uses low-confidence token selection with 128 denoising steps, a block length of 256, temperature $T = 0$, and classifier-free guidance (CFG) scale 0. For the sequence-to-sequence baselines, we use the official implementations. **Control-T5** uses the 4-control-token version with beam width 8, and **MUSS** uses its standard control preprocessors with beam width 5. Unless stated otherwise, other parameters follow the official defaults.

Random filling retains these settings but selects positions in random order. For remasking, we use the LLaDA implementation from Wang et al. (2025) and adapt it to Dream.

## B.2 LLM-as-a-Judge Inference Settings

Judge models use the prompt in Appendix A.2 to evaluate the candidate word forms from Section 2.3. Our primary judge is `DeepSeek-V4-Pro`, run with JSON output and maximum reasoning effort. Each candidate is evaluated once, and reported word forms outside the candidate list are discarded.

**System Prompt**

You are a conservative evaluator for English text simplification outputs. Your task is limited to identifying candidate word forms with orthographically malformed surface realizations that are consistent with subword-level generation or detokenization artifacts. Do not evaluate meaning preservation, grammaticality, fluency, or simplification quality. Do not flag valid paraphrases, well-formed words that are plausible in context, valid inflections, proper names, or technical terms. Flag a candidate only when there is concrete surface-form evidence for the label. Copy any reported word form exactly from the candidate list. Output only one valid JSON object, with no markdown fences or other text.

**User Prompt**

**Complex sentence:** `{complex_text}`
**Simplified references:** `{reference_1, . . . , reference_n}`
**Model output:** `{output}`
**Candidate word forms:** `{candidate_word_forms}`
The candidate word forms are selected from the model output by a novelty filter using whitespace splitting after stripping leading/trailing punctuation, and case-insensitive matching against the complex sentence and all simplified references.
Evaluate only the candidate word forms above. Assign a positive label only when a candidate's surface form is orthographically malformed and the evidence is consistent with a subword-level generation or detokenization artifact. This category includes invalid word-piece concatenation, missing whitespace between alphanumeric word pieces, overlapping duplication, prefix or suffix truncation, internal deletion, or internal substitution. Judge only the observable surface form; do not infer that the error was necessarily caused by a specific tokenizer or decoding mechanism.
The candidate list is a filtering heuristic: absence from the complex sentence or references is not, by itself, evidence of an error. Do not flag valid paraphrases, valid newly introduced words, valid inflections, valid proper names, valid technical terms, grammatical errors, semantic changes, or fluency problems. Do not flag a candidate that is a well-formed word and plausible in the local output context. A well-formed candidate that is a prefix, suffix, or substring of a longer source/reference word may be flagged as a fragment only when the local output context makes it implausible as a standalone word and the complex sentence or references support a specific longer intended word.
The reason must cite concrete surface-form evidence. Do not invent a spelling correction; claims about missing or extra characters are valid only when the intended corrected form differs from the candidate and is supported by the complex sentence, references, or local output context. The reason must not claim that a candidate should be corrected to the same string. If no candidate meets the criterion, set `subword_error` to `false` and `word_forms` to an empty list. If unsure, set `subword_error` to `false`. When reporting word forms, copy them exactly from the candidate list above.
Respond with exactly one JSON object with this shape: `{"subword_error": <bool>, "word_forms": <list of strings, possibly empty>, "reason": "<short English>"}`. The `word_forms` field must be a subset of the candidate list above.

Figure 5: LLM-as-a-Judge prompt used for detection.

| Evaluator | ASSET | | MED-EASi | |
|---|---|---|---|---|
| | # | % | # | % |
| Kimi-K2.6 | 18 | 5.01 | 30 | 10.00 |
| DeepSeek-V4-Pro | 19 | 5.29 | 33 | 11.00 |
| Human | 19 | 5.29 | 33 | 11.00 |

Table 7: Judge sensitivity on Dream-7B-Instruct unmasking outputs.

Table 7 compares `DeepSeek-V4-Pro` with an alternative judge, `Kimi-K2.6`, on Dream-7B-Instruct unmasking outputs. `DeepSeek-V4-Pro` matches the human-verified counts, while `Kimi-K2.6` reports fewer MED-EASi SBCs, so we use `DeepSeek-V4-Pro` in the main analyses.

## C LLaDA-1.5 Decoding Ablations

Tables 8 and 9 mirror the Dream ablations in Section 4.2. LLaDA-1.5 shows the same pattern as Dream, with higher parallelism substantially increasing SBC rates (from 0.00% to 20.67% on MED-EASi), while varying the denoising budget at parallelism 2 does not consistently reduce SBC rates (fluctuating between 1.67% and 6.00% on MED-EASi). This suggests that boundary conflicts are driven more by parallel token commitment than by an insufficient step budget.

| Parallelism | Steps | ASSET Test (359) | MED-EASi Test (300) |
|---|---|---|---|
| 1 | 128 | 0 (0.00) | 0 (0.00) |
| 2 | 128 | 5 (1.39) | 8 (2.67) |
| 4 | 128 | 30 (8.36) | 30 (10.00) |
| 8 | 128 | 70 (19.50) | 62 (20.67) |

Table 8: LLaDA-1.5 SBC rates by parallelism under standard unmasking (128 steps).

| Parallelism | Steps | ASSET Test (359) | MED-EASi Test (300) |
|---|---|---|---|
| 2 | 32 | 9 (2.51) | 18 (6.00) |
| 2 | 64 | 6 (1.67) | 5 (1.67) |
| 2 | 128 | 5 (1.39) | 8 (2.67) |
| 2 | 256 | 3 (0.84) | 7 (2.33) |
| 2 | 512 | 3 (0.84) | 10 (3.33) |

Table 9: LLaDA-1.5 SBC rates by denoising steps under standard unmasking (parallelism 2).

# D Generality Across Tasks and Architectures

## D.1 Grammatical Error Correction

We evaluate Dream-7B-Instruct and LLaDA-1.5 on all 747 JFLEG test sentences (Napoles et al., 2017), using each model's default unmasking schedule with parallelism 2. Table 10 reports GLEU, including published comparison scores from Coyne et al. (2023), while Table 11 reports the number of outputs containing a human-verified SBC. Both models produce nonzero SBC rates while substantially outperforming the uncorrected source in GLEU. Representative cases appear in Table 12.

| Model | JFLEG Test GLEU |
|---|---|
| Source (Uncorrected) | 40.54 |
| Human References | 62.37 |
| ELECTRA-VERNet | 61.61 |
| Dream-7B-Instruct | 61.12 |
| LLaDA-1.5 | 58.77 |

Table 10: GLEU scores on the JFLEG test set.

| Model | Method | Parallelism | JFLEG Test (747) |
|---|---|---|---|
| Dream-7B-Instruct | Unmasking | 2 | 7 (0.94%) |
| LLaDA-1.5 | Unmasking | 2 | 7 (0.94%) |

Table 11: SBC rates on the JFLEG test set.

| Model | Role | Sentence |
|---|---|---|
| Dream-7B-Instruct | Source | Adult content to sent their chirdren to school to contact these different kinds of class. |
| | Reference | Adults are content to send their children to school to experience these different kinds of classes. |
| | Output | AdultAdults send their children to school to contact these different kinds of classes. |
| LLaDA-1.5 | Source | traditional enclopedias are written by trained expert for academic rigor that can not really achieve. |
| | Reference | Traditional encyclopedias are written by trained experts for academic rigor that can not really achieve. |
| | Output | Traditional encyclopedias are written by trained experts for academic rigoror, which they cannot truly achieve. |

Table 12: Representative SBC cases from JFLEG grammatical error correction. Red highlights malformed word forms.

## D.2 Block-Diffusion Model

Fast-dLLM v2-7B generates blocks autoregressively while refining tokens in parallel within the block (Wu et al., 2026a). We fix the block size at 32 and confidence threshold at 0.9, then vary the sub-block size over {4, 8, 16, 32}. Table 13 shows that SBC rates increase with sub-block size on ASSET and rise before plateauing on MED-EASi, but even the smallest setting does not eliminate the errors. This extends the observed relationship between parallel token commitment and SBC prevalence to a block-diffusion architecture.

| Block Size | Sub-block Size | ASSET (359) | MED-EASi (300) |
|---|---|---|---|
| 32 | 4 | 22 (6.13%) | 30 (10.00%) |
| 32 | 8 | 29 (8.08%) | 39 (13.00%) |
| 32 | 16 | 39 (10.86%) | 40 (13.33%) |
| 32 | 32 | 49 (13.65%) | 40 (13.33%) |

Table 13: Human-verified SBC counts and rates for Fast-dLLM v2-7B on the ASSET (359) and MED-EASi (300) test sets.

# E Case Studies

## E.1 Taxonomy Cases

Table 14 shows ASSET unmasking examples for the six taxonomy categories in Section 2.2.

## E.2 LLM-as-a-Judge Cases

Figure 6 shows a representative judge JSON output for a T6 *Invalid Concatenation Error*, where *Australian* is fused with the *ways* suffix from *Railways*.

## E.3 Non-DLM Cases

Table 15 shows representative examples from non-DLM systems. These errors appear in only a few outputs, unlike the repeated decoding-related SBC pattern in DLMs.

## E.4 Attention Cases

We use the final layer as a representative attention view. Figure 7 shows this shifted-logit view for the *Australianways* case discussed in Section 4.5, focusing on the relevant source and output tokens.

## E.5 Regeneration Cases

Table 16 presents representative unmasking regeneration cases across DLMs and datasets, illustrating how local regeneration replaces malformed output spans caused by subword boundary conflicts in the initial model outputs.

| Category | Source | Sentence |
|---|---|---|
| **T1:** Overlapping Duplication Error | Complex | Perry Saturn (with Terri) defeated Eddie Guerrero (with Chyna) to win the WWF European Championship (8:10) Saturn pinned Guerrero after a Diving elbow drop. |
| | Output | Perry Saturn defeated Terri and Eddie Guerrero with Chynayna to win the WWF European Championship. |
| **T2:** Prefix Truncation Error | Complex | In architectural decoration Small pieces of colored and iridescent shell have been used to create mosaics and inlays, which have been used to decorate walls, furniture and boxes. |
| | Output | ored shell has been used to decorate walls, furniture, and boxes. |
| **T3:** Suffix Truncation Error | Complex | The Amazon Basin is the part of South America drained by the Amazon River and its tributaries. |
| | Output | The Amazon Basin is the area in South America where the Amazon River and its trib flow. |
| **T4:** Internal Deletion Error | Complex | It was discovered by Stephen P. Synnott in images from the Voyager 1 space probe taken on March 5, 1979 while orbiting around Jupiter. |
| | Output | Stehen P. Synnott found it in images from the Voyager 1 space probe while it was around Jupiter in March 1979. |
| **T5:** Internal Substitution Error | Complex | The Vezina Trophy is awarded annually to the National Hockey League's goaltender who is "adjudged to be the best at this position". |
| | Output | The Venderina Trophy is given every year to the goaltender who is the best in the NHL. |
| **T6:** Invalid Concatenation Error | Complex | Alessandro ("Sandro") Mazzola (born 8 November 1942) is an Italian former football player. |
| | Output | Alessandro Mazzola, also knownSandro, is an Italian former football player. He was born on November 8, 1942. |

Table 14: Representative Subword Boundary Conflict examples for the six taxonomy categories. Yellow marks fragments preserved between Complex and Output, red marks fragments dropped in the Output, and blue marks fragments inserted in the Output.

| Model | Source | Sentence |
|---|---|---|
| **LLaMA-3-8B-Instruct** | Complex | The Old Testament's examples of killings sanctioned by God are often cited in defense of the view that "murder" is a more accurate translation. |
| | Output | The Old Testament shows examples of people killing others with God's permission. Some people use these examples to argue that'murder' is a better translation of the word. |
| **LLaMA-3-8B-Instruct** | Complex | The word is derived from 'pahar' meaning mountain. |
| | Output | The word comes from the Hindi word 'pahar', which means'mountain'. |
| **Qwen2.5-7B-Instruct** | Complex | It was originally thought that the debris thrown up by the collision filled in the smaller craters. |
| | Output | ly, people believed that the debris from the collision filled in the smaller craters. |
| **Qwen2.5-7B-Instruct** | Complex | This resulted in a vest with vertical stripes of white and navy-blue color. |
| | Output | ed in a vest with vertical white and navy-blue stripes. |
| **Control-T5** | Complex | Bagnotar is one of the 51 Union Councils of Abbottabad District in the North-West Frontier Province of Pakistan. |
| | Output | Bagnotar is one of the 51 Union Councils of Abbottabad District in the North-West Frontier Province of PakistanAdministrative. |
| **Control-T5** | Complex | Patients with unexplained or sustained worsening of symptoms should be evaluated for pancreatic cancer, particularly if assessment reveals a pancreatic duct stricture. |
| | Output | Patients with unexplained or worsening of symptoms should be looked at for pancreatic cancer. This may be because of a pancreatic duct tighture. |

Table 15: Representative non-DLM corner cases illustrating that SBC-like surface forms are not exclusive to DLMs. Yellow marks fragments preserved between Complex and Output, red marks fragments dropped in the Output, and blue marks fragments inserted in the Output.

| Model | Dataset | Source | Sentence |
|---|---|---|---|
| Dream-7B-Instruct | ASSET | Complex | Before 1980, a section of the Commonwealth Railways Central Australian line passed along the western side of the Simpson Desert. |
| | | Output | Before 1980, a part of the Central Australianways line went through the Simpson Desert. |
| | | Regeneration | Before 1980, a part of the Central Australian Railways line went through the Simpson Desert. |
| | MED-EASi | Complex | In the most severe form (called malignant yellow fever), delirium, intractable hiccups, seizures, coma, and multiple organ failure may occur terminally. |
| | | Output | In the most severe form of yellow fever, peopleirium, hiccups hiccups, seizures, coma, and multiple organ failure may occur. |
| | | Regeneration | In the most severe form of yellow fever, people may become confused, hiccups hiccups, seizures, coma, and multiple organ failure may occur. |
| LLaDA-1.5 | ASSET | Complex | The four canonical texts are the Gospel of Matthew, Gospel of Mark, Gospel of Luke and Gospel of John, probably written between AD 65 and 100. |
| | | Output | The four main Gospelospels are Matthew, Mark, Luke, and John. |
| | | Regeneration | The four main Christian gospels are Matthew, Mark, Luke, and John. |
| | MED-EASi | Complex | Intractable apneic spells require ventilator support. |
| | | Output | Unicult breathing problems need breathing support. |
| | | Regeneration | People who have trouble breathing need breathing support. |
| LLaDA-8B-Instruct | ASSET | Complex | At four-and-a-half years old he was left to fend for himself on the streets of northern Italy for the next four years, living in various orphanages and roving through towns with groups of other homeless children. |
| | | Output | A four-year-old boy lived in northernages and Italian orphanages with other homeless children. |
| | | Regeneration | He spent four years on the streets in northern Italy, and lived in orphanages with other homeless children. |
| | MED-EASi | Complex | Spontaneous remissions may last years. |
| | | Output | Somes of remission can last for years. |
| | | Regeneration | Spontaneous remissions can last for several years. |

Table 16: Representative unmasking regenerations, with one example for each model and dataset combination. Red marks malformed output spans, and blue marks regenerated corrections.

**LLM-as-a-Judge Json Output**


```
{
  "id": "asset_test_000325",
  "line": 325,
  "complex": "Before 1980, a section of the Commonwealth Railways Central
      Australian line passed along the western side of the Simpson Desert.",
  "simple_reference": "In Australia prior to 1980 a part of the Central
      Australian railway passed along the Simpson Desert.",
  "output": "Before 1980, a part of the Central Australianways line went
      through the Simpson Desert.",
  "candidate_word_forms": ["Australianways", "through"],
  "judge": {
      "subword_error": true,
      "word_forms": ["Australianways"],
      "reason": "Orthographically malformed non-word resulting from
      concatenation of 'Australian' and 'ways' without an intervening space."
  }
}
```


Figure 6: Representative LLM-as-a-Judge JSON output. Long strings are wrapped for readability; metadata fields (`corpus`, `split`) are omitted for space.

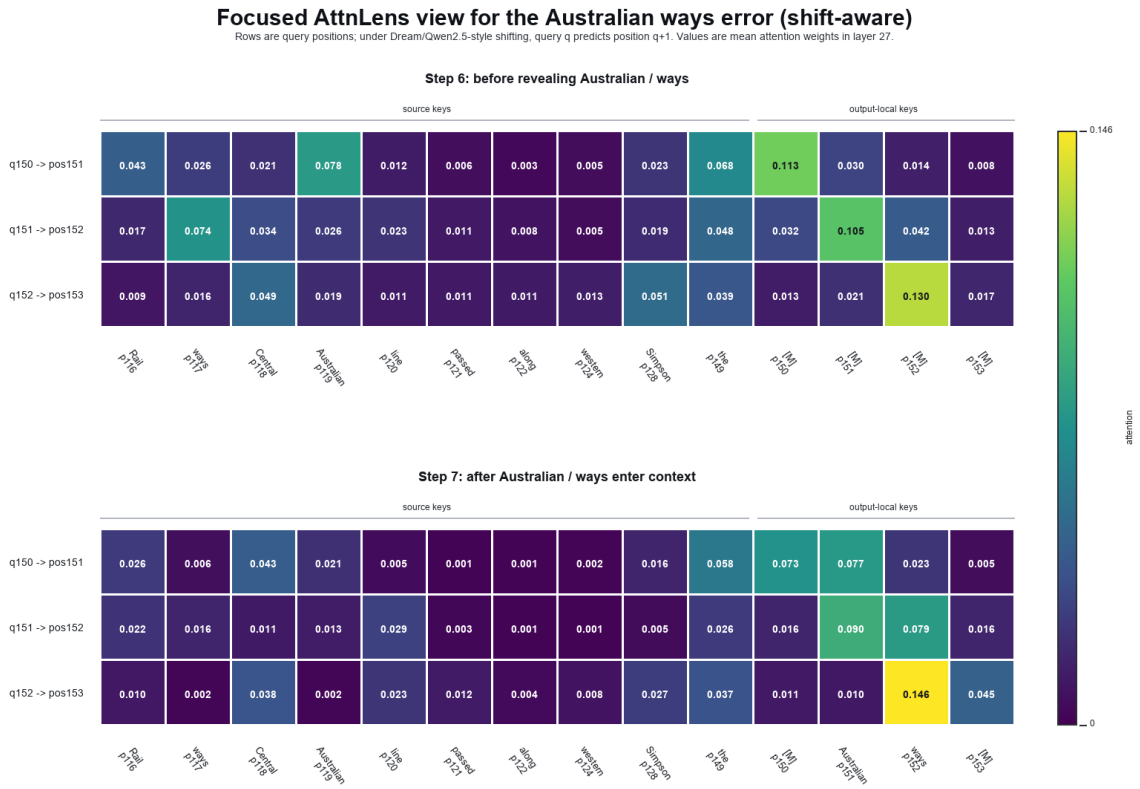



Figure 7: Focused attention view for the Dream-7B-Instruct *Australianways* case. Values are final-layer (layer-27) mean attention weights; rows are interpreted with Dream's shifted logits.